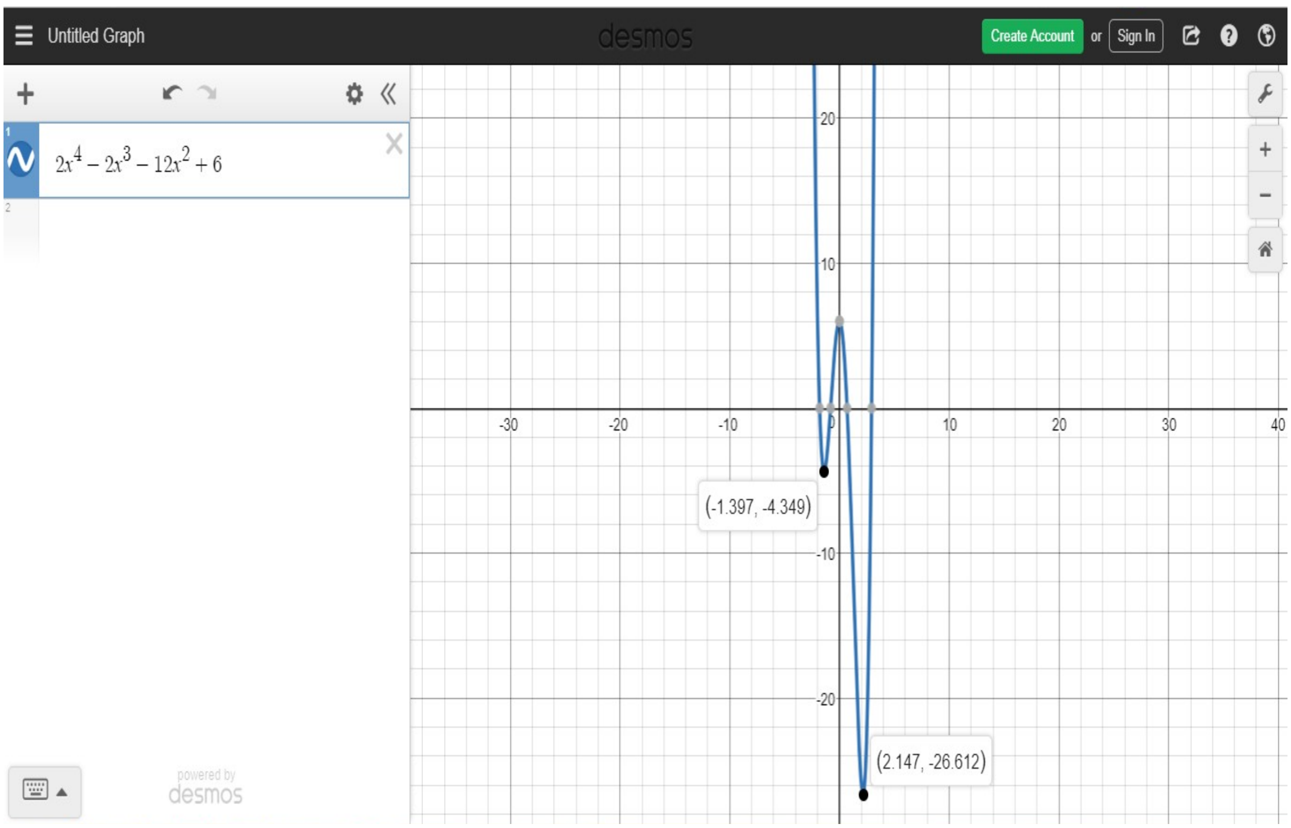
MACHINE LEARNING
ASSIGNMENT-4
WRITTEN EXERCIZE

Submitted by:
Nishika Chopra nc2259 N18598308
Karanpreet Sign Wadhwa ksw352 N13337853

Answer 1:
a) Value of x at-

Local minima: -1.397

Global minima: 2.147



b) Values of x and f(x) for 6 iterations at:

Start: -4,-512
Iteration 1: -3.488,-328.769
Iteration 2: -3.15923,-236.315
Iteration 3: -2.92292,-180.885
Iteration 4: -2.74203,-144.237
Iteration 5: -2.5978,-118.395
Iteration 6: -2.4794,-99.3144

Last 6 values of x and f(x) for 1200 iterations:

Iteration 1: -1.39718,-5.34058e-05
Iteration 2: -1.39718,-5.34058e-05
Iteration 3: -1.39718,-5.34058e-05
Iteration 4: -1.39718,-5.34058e-05
Iteration 5: -1.39718,-5.34058e-05
Iteration 6: -1.39718,-5.34058e-05
Yes, the value of x has converged to a minimum value of x=-1.39718. This is the local minima.

c) Values of x and f(x) for 6 iterations at:

Start: 4,320
Iteration 1: 3.68, 229.114
Iteration 2: 3.45089, 174.489
Iteration 3: 3.2764, 138.329
Iteration 4: 3.13807, 112.818
Iteration 5: 3.02525, 93.9812
Iteration 6: 2.93127, 79.5871

Last 6 values of x and f(x) for 1200 iterations:
Iteration 1: 2.14718, 0.000106812
Iteration 2: 2.14718, 0.000106812
Iteration 3: 2.14718, 0.000106812
Iteration 4: 2.14718, 0.000106812
Iteration 5: 2.14718, 0.000106812
Iteration 6: 2.14718, 0.000106812

Yes, the value of x has converged to a minimum value of x=2.14718. This is the global minima.

d) Values of x and f(x) for 6 iterations at:
(x=-4 and n=0.01)
Start: -4,-512
Iteration 1: 1.12,-23.167
Iteration 2: 1.35167,-23.646
Iteration 3: 1.58813,-21.204
Iteration 4: 1.80017,-15.9785
Iteration 5: 1.95995,-9.85533
Iteration 6: 2.05851,-5.04623

Last 6 values of x and f(x) for 1200 iterations:
(x=-4 and n=0.01)
Iteration 1: 2.14718,-3.8147e-06

Iteration 2: 2.14718,-3.8147e-06
Iteration 3: 2.14718,-3.8147e-06
Iteration 4: 2.14718,-3.8147e-06
Iteration 5: 2.14718,-3.8147e-06
Iteration 6: 2.14718,-3.8147e-06

When we compare these values to the ones obtained by taking value of learning factor as 0.001 we find that in the first case the minimum value of x was found at local minima while in this case the value of x is found at global minima. This happened due to the fact that we increased the value of the learning factor and it skipped the local minima and hence converged at global minima.

e) Values of x and f(x)when n=0.1:

Start:-4,-512
Iteration 1: 47.2,826733
Iteration 2: -82626.1,-4.51279e+15
Iteration 3: 4.51279e+14,inf
Iteration 4: -inf,-nan
Iteration 5: -nan,-nan
Iteration 6-100: -nan,-nan
Due to very high learning factor the minima was skipped and thus we could not achieve any minimum value for x.

Answer 2)

a) In stochastic gradient we update weights after each sample. Thus, after 100 epochs for 500 samples we will have to update 500x100= 50000 times.
b)

ans 2) (i) $\Delta v_{23} = -\eta \dfrac{\partial E(w, v|x)}{\partial v_{23}}$

$= -\eta \dfrac{\partial}{\partial v_{23}} \left[\dfrac{1}{2} \left\{ 3(r_1-y_1)^2 + 7(r_2-y_2)^2 \right\} \right]$

$= -\eta \left[\dfrac{\partial}{\partial v_{23}} \left[\dfrac{1}{2} \left\{ 3(r_1-y_1)^2 \right\} \right] + \dfrac{\partial}{\partial v_{23}} \left[\dfrac{1}{2} \left\{ 7(r_2-y_2)^2 \right\} \right] \right]$

$= -\eta \left[3(r_1-y_1)\left(-\dfrac{\partial y_1}{\partial v_{23}}\right) + 7(r_2-y_2)\left(-\dfrac{\partial y_2}{\partial v_{23}}\right) \right]$

$= -\eta \left[3(r_1-y_1) \times 0 + 7(r_2-y_2)(-z_3) \right]$

$= \eta \times 7(r_2-y_2)(z_3)$

Ans.

ans 2) ~~(iii)~~ (ii) $\Delta W_{hj} = -\eta \frac{\partial E}{\partial w_{hj}}$

$$\Delta W_{hj} = -\eta \sum \frac{\partial E^t}{\partial y^t}\cdot\frac{\partial y^t}{\partial z_h^t}\cdot\frac{\partial z_h^t}{\partial w_{hj}}$$

$$= -\eta \sum \underbrace{\left[-3(r_1-y_1)-7(r_2-y_2)\right]}_{\partial E^t/\partial y^t}\cdot \underbrace{v_h}_{\partial y^t/\partial z_h^t}\cdot \underbrace{z_h^t(1-z_h^t)x_j^t}_{\partial z_h^t/\partial w_{hj}}$$

$$= \eta \sum \left[3(r_1-y_1)+7(r_2-y_2)\right] v_h\, z_h^t(1-z_h^t)\, x_j^t$$

$$= \eta \sum \left[3(r_1-y_1)v_h z_h^t(1-z_h^t)x_j^t + 7(r_2-y_2)v_h z_h^t(1-z_h^t)x_j^t\right]$$

Ans.

Note :⇒ $\frac{\partial E^t}{\partial y^t} = \frac{\partial}{\partial y^t}\left\{\frac{1}{2}\left[3(r_1-y_1)^2+7(r_2-y_2)^2\right]\right\}$

$$= \frac{\cancel{2}}{\cancel{2}}\left[3(r_1-y_1)(-1)+7(r_2-y_2)(-1)\right]$$

$$= -3(r_1-y_1) - 7(r_2-y_2)$$

Answer 3)

a) NeuralNetRK
For negative values we will use neural NeuralNetRK since it will take negative and real values.

b) NeuralNetCK
It is given that: sum of the yi is 1. Therefore, for K outputs the sum of all the outputs will be equal to 1. This is known as softmax which gives estimated probabilities for the K labels of x.

c) NeuralNetCK
In this problem, there are multiple classes namely face, cat and tree. There are three output values p1,p2 and p3 which are probabilities of the image being a face,cat or tree respectively. Thus according to the definition NeuralNetCK would be the most appropriate in this case because it is used to classify when there are more than 2 classes (K>2). Moreover, it gives the probability of an example belonging to a class and the sum of the outputs has to be 1.

d) NeuralNetRZeroOne
In this text classification problem, there are K outputs in the range of {0,1}. Each output is an element in the set {0,1} and thus it will have some real values. Also the output is either 0 or 1. Thus NeuralNetRZeroOne suits this perfectly.

Answer 4)

a) In Random forests, trees are created with different attributes and the values encoded do not matter. Different trees with their roots as attributes are created and the tree with highest efficiency is chosen. Contrastingly, in neural nets the product of input and weights are taken. Therefore if we assign random weights, the product will be wrong, and similarly the output will also come out to be wrong. Hence, it would be fine to do label encoding if we were using a random forest, rather than a neural net.

b) (i)

ans 4) (b) (i) We would replace the stalk shape attribute with 2 attributes called S_1 and S_2, where iff stalkshape = tapering, $S_1 = 1$ and iff stalk shape = enlarging, $S_2 = 1$

Transformed dataset :-

	Z_1	Z_2	Z_3	Z_4	S_1	S_2	Label
$x^{(1)}$	0	0	0	1	1	0	0
$x^{(2)}$	0	0	1	0	0	1	0

(ii) An ordinal attribute whose values are low, medium, and high are relational and they have a particular order between them. For example it the attribute is temperature whose values are low, medium and high, these values are interrelated to each other and thus can be encoded.

Low<Medium<High

Hence we can assign them values as 1,2 and 3.

We apply one hot encoding when the values that are close to each other in the label encoding correspond to target values that aren't close (non - linear data). Whereas we apply label encoding when we can come up with a label encoder that assigns close labels to similar categories: This leads to less splits in the tress hence reducing the execution time.

(iii) For nominal attributes with only two values, it's generally fine to just represent the two values as 0 and 1 (or -1 and +1), rather than using one-hot encoding as there are can only be two values for the attribute. If one is true the other will be false and vice versa. The attribute can have only two values and thus there is no point of using one hot encoding.